DOCUMENT RESUME

ED 320 943 TM 015 206

AUTHOR Friedman, Charles B.
TITLE Critical Items and Standard Setting for Health Certification Examinations.
PUB DATE Mar 89
NOTE 12p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, March 28-30, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Cutting Scores; *Evaluators; *Health Occupations; Higher Education; *Licensing Examinations (Professions); National Programs; *Performance; *Scoring; Test Construction; Testing Programs; *Test Items
IDENTIFIERS *Standard Setting

ABSTRACT

In most judgmental methods for setting performance standards, content experts evaluate test items based on a conceptualization of a borderline examinee or a group of examinees. An alternative method is proposed by which judges evaluate items based on the degree of criticality regardless of content area. This classification of items serves to link the job analysis stage of test development to the standard setting process of examination administration. Based on the ratings of seven judges of a national health certification examination administered in March of 1988 to 635 entry-level examinees, a special critical item section score was calculated. This subscore was evaluated for its internal consistency. Examinee performance on critical items was compared to the total test score and evaluated for its potential impact on the number of candidates passing or failing the test and as a criterion in setting a standard score. Implications for test validity are discussed. Three data tables are included. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

CHARLES FRIEDMAN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED320943

CRITICAL ITEMS

AND

STANDARD SETTING FOR HEALTH CERTIFICATION EXAMINATIONS

Charles B. Friedman

The Psychological Corporation

Paper presented at the Annual Meeting of The National Council on Measurement in Education

San Francisco, CA

March 1989

TM015206

BEST COPY AVAILABLE

ABSTRACT

In most judgmental methods for setting performance standards, content experts evaluate test items based on a conceptualization of a borderline examinee or a group of examinees. This study proposes an alternative method where judges evaluate items based on the degree of criticality regardless of content area. This classification of items will serve to link the job analysis stage of test development to the standard setting process of examination administration.

Based on judges' ratings of a national health certification examination, a special critical item section score was calculated. This subscore was evaluated for its internal consistency. Examinee performance on critical items was compared to the total test score and evaluated for its potential impact on the number of candidates passing/failing the test and as a criterion in setting a standard score. Implications for test validity are discussed.

CRITICAL ITEMS AND STANDARD SETTING FOR HEALTH CERTIFICATION EXAMINATIONS

Setting the standard or minimum passing score for credentialing examinations remains a difficult but crucial process. Over 35 methods for setting or adjusting standards have been identified (Berk, 1986); almost all require experts to make judgments. Commonly used judgmental score setting methods (Angoff, Nedelsky, Ebel) require judges to conceptualize a borderline candidate or a group of individuals and evaluate their projected performance on an item or distractor (Meskauskas, 1986). An alternative method would focus on the item and the consequence of incorrectly answering the item; i.e., would an incorrect response to an item have a critical effect on the patient? While this does involve a conceptualization by judges, it is a conceptualization of an item and the potential effect of the item on the consumer rather than the conceptualization of a hypothetical examinee, hypothetical group of examinees, or known group of examinees.

The term criticality or critical is not a foreign one in the developmental process of licensing and certification examinations. On the contrary, it is an essential ingredient in the early phases of test construction. The foundation of licensing and certification examination is the task or job analysis: How important or critical is the successful performance of a task? The answer to this question as well as the frequency of occurrence sets the foundation for the other steps in the test development process and more importantly is cited as an essential element in the content validity of the test.

Criticality also plays an important part in the development of test specifications; a number of methods have been suggested to combine criticality, frequency of occurrence, and judgmental weightings in the

building of specifications (Schmeiser, 1988). However, the role of criticality generally has ended at this phase of the test development process.

If one were to conceptualize the test building process as a series of steps, we could represent the early phases of test development as follows:

Step 1: Job and Task Analyses
Step 2: Test Specifications/Blueprint
Step 3: Item Writing
Step 4: Form Assembly

The establishment of content validity, that the test items are representative of the jobs and tasks (their frequency and criticality), relies on the links between Step 1: job analysis and Step 2: test blueprint and then Step 3: the writing of test items.

1 → 2
2 → 3
1 → 3

An underlying assumption is that criticality which is present at Step 1 will also be present at Steps 3 based on the transitive link of test specifications. There is, however, no direct path between Step 1 and Step 3 in either direction. Moreover, the methods for developing job descriptions and translating the descriptions into test specifications are not well developed. (Kane, Kingsbury, Colton, and Estes, 1989). The presence of a stronger link between the job analysis, and actual test items, and other steps in the test construction process would present additional validation evidence.

The purpose of this explanatory study is thus twofold in nature:

1. Examine the characteristics of a critical item section score and its effect on the passing rate of examinees.

2. Suggest how the examinee performance on critical items can serve to link the steps of test development and provide additional evidence for construct validity.

Instrument

A special Item Criticality Questionnaire was developed to rate each test question. The original instrument was based on a four-point integer scale suggested by Schmeiser (1988); discussion with a committee of subject experts redefined criticality as surgical consequence and reduced this to a three-point integer scale because of the nature of the health occupation. It was felt by the experts that the continuum for this profession was direct surgical consequence, indirect surgical consequence, no surgical consequence. Directions on the rating instrument were as follows:

Read each item on the enclosed test.

For each item on the test, answer the following question:

If an examinee answered this question incorrectly, what would be the potential surgical consequence on a patient's health?

- Direct surgical consequence affecting the patient's health
- Indirect surgical consequence affecting the patient's health
- No surgical consequence and no inconvenience to patient

Indicate your response on the following pages by placing a check in the appropriate column.

The above rating instrument was applied to a previously administered 250 multiple-choice item examination. This examination was constructed from a test blueprint based on an incumbent job analysis that focused on the criticality and importance of tasks.

Sample

A sample of seven content experts responsible for examination development were asked to complete the Item Criticality Questionnaire. These item raters included both educators and practitioners.

The empirical data used in the study was based on the March 1988 administration of the health certification examination to 635 entry-level examinees across the country.

Method

Item raters discussed the meaning of direct, indirect, and no surgical consequence at an examination committee meeting. Each rater was then sent a packet by mail consisting of the March 1988 test booklet, instructions and categories for rating items, and a form on which to record the response to each test item.

The analysis used to address the first objective of this study was based on the modal rating to each item of the rater's response. Because of the 1-3 integer nature of the scale and the small number of judges in the sample, the mode was selected as the best measure of central tendency and the criterion to classify each item into a criticality category. Items that were bimodal--i.e., three judges classifying an item as direct consequence, three indirect consequence, and one no consequence--were not classified.

Items classified as direct surgical consequence were then grouped together to form a Critical Item Section score from the March 1988 test administration. The mean, standard deviation, range, and Kuder-Richardson reliability of this section were then calculated.

Angoff standard setting procedures had been used to establish a passing score based on the total test for the March 1988 examination. Using the individual item Angoff rated data, a new passing score based only on the Critical Item Section was calculated. The number of examinees who would have passed the test based on the Critical Item Section cut score was then compared to the number of examinees who passed the test based on the total test score.

Results and Discussion

Table 1 provides item classifications based on judges' response to 200 test items. Because incomplete information was available for the last 50 items, they were eliminated from the analysis. Forty-two items were classified as having a direct surgical consequence; i.e., an incorrect response could result in a direct surgical consequence affecting a patient's health. The largest number of items, 72, were placed in the middle category, indirect surgical consequence. Sixty-seven items were classified as no surgical consequence and 19 items were unclassified. Of the 42 items classified as direct consequence, 25 were placed in this category based on the agreement of five or more of the raters.

Table 2 presents the mean, standard deviation, range, and reliability for the Critical Item Section. The mean score of the 42 items was 34.7; the reliability (KR-20) was .64. This compares to a reliability of .92 based on the total test score.

The data suggest that raters can reach some agreement in classifying items based on their criticality or degree of surgical consequence and how an incorrect response to an item might affect a patient. While the Critical Item Section reliability was lower than the total test reliability, this was

affected by both a smaller number of questions and the fact that examinees generally performed better on the critical items than the test as a whole. The major focus of the first objective of this study is a comparison of the examinee passing rate based on the total test score and the Critical Item Section score. Table 3 presents information on the number of individuals passing and failing the test based on these two criteria.

Of 635 candidates taking the examination, 493 passed based on the total test score; 609 would have passed based on the Critical Item Section score. It is interesting to note that of the 493 candidates passing the total test, only 1 candidate would have failed had the passing score been based only on the Critical Item Section. Thus, it appears that those passing based on the total test are also passing the critical items, those items which could potentially cause the most harm to the patient or consumer.

The above suggests that using the Critical Item Section score along with the total test score as criteria for passing this examination would have provided very little additional useful information in evaluation examinee performance; use of the total test score as criterion was sufficient. Because the critical items provide very little additional information, it also suggests that the total examination is doing a good job in measuring the critical skills and tasks needed by a minimally competent individual.

However, another scenario may suggest a different conclusion. In this scenario, 493 individuals passed the test based on the total test score. Of these 493, 100 received failing scores based on the Critical Item Section. This would suggest that 100 candidates who could potentially cause great harm to a surgical patient would have passed the test. This would create not only

a dangerous situation for the patient, but also a tenuous one for the test constructor.

It would also suggest that the actual test items and standard setting procedures are not related to the original job analysis and test specifications. The influence and role of the criticality factor in the task analysis became lost or obscured during the latter stages of test development. The implied transitive relationship in the test building process may have been faulty and the validity of the test questionable.

Too often the test construction process is one of discrete steps that flows in one direction.

Job Analysis
↓
Test Blueprint
↓
Item Writing
↓
Test Assembly
↓
Standard Setting

The above generalized version of the test blueprint only proceeds in one direction. It is assumed that at each step the careful rigorous procedures by themselves are sufficient evidence for the construct validity of the test. However, the links between the steps may be weak and efforts must be made, not only to strengthen the links, but also to create a flow or check that connects steps with earlier steps in the development process.

This paper has tried to suggest the application of a standard setting procedure based on critical items as an additional means to provide evidence of construct validity.

Table 1

Classification of Items by Raters

Category	Number of Items
Direct Surgical Consequence	42
Indirect Surgical Consequence	72
No Surgical Consequence	67
Unclassified	19

Table 2

Critical Item Section Data

Number of Items	42
Mean Score	34.7
Standard Deviation	3.7
Range	20-42
Kuder-Richardson 20	.64

N = 635

Table 3

Number of Candidates Passing/Failing
Based on the Total Test and Critical Item Section

Based on Total Test	Based on Critical Item Section		
	Pass	Fail	Total
Pass	492	1	493
Fail	117	25	142
Total	609	26	635

References

Berk, R. A. (1986). A Consumer's Guide to Setting Performance Standard on Criterion-Referenced Tests. Review of Educational Research, 56, 137-1

Kane, M., Kingsburg, C., Colton, D. & Estes, C. (1989) Combining Data on Criticality and Frequency in Developing Test Plans. Journal of Educational Measurement, 26, 17-28.

Meskauskas, J. A. (1986). Setting Standard for Credentialing Examinations An Update. Evaluation and the Health Professions, 9, 187-204.

Schmeiser, C. B. (1988, April). Effects of Statistical and Judgemental Weightings of Task Analysis Data on Test Specifications. Paper presen at the Annual Meeting of the National Council on Measurement in Education. New Orleans, LA.

8595I